PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04

A1

(11) International Publication Number: **WO 99/39004**

(43) International Publication Date: 5 August 1999 (05.08.99)

(21) International Application Number: PCT/US98/05438

(22) International Filing Date: 19 March 1998 (19.03.98)

(30) Priority Data:

60/073,345	2 February 1998 (02.02.98)	US
60/073,853	2 February 1998 (02.02.98)	US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application

US	60/073,345 (CIP)
Filed on	2 February 1998 (02.02.98)

(71) Applicants *(for all designated States except US)*: AFFYMETRIX, INC. [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US). THE GOVERNMENT OF THE UNITED STATES OF AMERICA, SECRETARY, DEPARTMENT OF HEALTH AND HUMAN SERVICES [US/US]; National Institutes of Health, Patent Branch, Office of Technology Transfer, Suite 325, 6011 Executive Boulevard, Rockville, MD 20850 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: CHEE, Mark [AU/US]; 3199 Waverly Avenue, Palo Alto, CA 94306 (US). HACIA, Joseph, G. [US/US]; 263 Congressional Lane, Rockville, MD 20852 (US). COLLINS, Francis, S. [US/US]; 5908 Tudor Lane, Rockville, MD 20852 (US). EDGEMON, Keith [-/US]; National Institutes of Health, Building 49, Room 3A14, 49 Convent Drive, MSC 4442, Bethesda, MD 20892-4442 (US).

(74) Agents: LIEBESCHUETZ, Joe et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th floor, San Francisco, CA 94111-3834 (US).

(81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: ITERATIVE RESEQUENCING

(57) Abstract

The invention provides iterative methods of analyzing a target nucleic acid that represents a variant of a reference nucleic acid. An array of probes is designed to be complementary to an estimated sequence of a target nucleic acid. The array of probes is then hybridized to the target nucleic acid. The target sequence is reestimated from hybridization pattern of the array to the target nucleic acid. A further array of probes is then designed to be complementary to the reestimated sequence, and this array is used to obtain a further reestimate of the sequence of the target nucleic acid. By performing iterative cycles of array design and target sequence estimation, the estimated sequence of the target converges with the true sequence.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

ITERATIVE RESEQUENCING

CROSS-REFERENCE TO RELATED APPLICATIONS

This application derives priority from USSN 60/073,345, filed February 2, 1998, and Townsend & Townsend & Crew Attorney Docket No. 018547-030510US, filed February 2, 1998, each of which is incorporated by reference in its entirety for all purposes.

ACKNOWLEDGMENT OF GOVERNMENT SUPPORT

This work was financed in part by grant number 5POLHGO1323-03.

TECHNICAL FIELD

The invention resides in the technical fields of molecular genetics, genomics and comparative sequence analysis.

BACKGROUND

The traditional approach to genome sequence analysis requires a primary sequence to be determined by conventional gel-based methods (typically using Applied Biosystems DNA sequencers). In this type of approach, the amount of work increases in proportion to both the length of sequence and the number of organisms tested and becomes impractical for large stretches of DNA or large numbers of organisms. For this reason, relatively few individuals within a species have been sequenced to look for polymorphic variation. Furthermore, only a few exemplary species, such as humans and *E. coli*, have been subject to large-scale sequencing.

Arrays of probes provide a more efficient means of analyzing variant sequences once a prototypical or reference sequence has been determined. Analysis of the hybridization pattern of probes to a target nucleic acid reveals the position, and optionally the nature, of differences between the target and reference sequence. For example, WO 95/11995 describes arrays comprising four probe sets. Comparison of the intensities of four corresponding probes from the four sets to a target sequence reveals the identity of a corresponding nucleotide in the target sequences aligned with an interrogation position of the probes. The corresponding nucleotide is the complement of the nucleotide occupying the interrogation position of the probe showing the highest intensity.

The existence of variation between a target and reference sequence can also be identified by differences in normalized hybridization intensities of probes flanking the variation when the probes are respectively hybridized to target and reference sequences. Relative loss of hybridization intensity is manifested as a "footprint" of probes flanking the point of variation between target and reference sequence (see EP 717,113, incorporated by reference in its entirety for all purposes). Additionally, hybridization intensities for multiple targets from different sources can be classified into groups or clusters suggested by the data, not defined *a priori*, such that isolates in a give cluster tend to be similar and isolates in different clusters tend to be dissimilar (see WO 97/29212, incorporated by reference in its entirety for all purposes).

Array-based resequencing has been used, for example, in the identification of large numbers of human polymorphisms in mitochondrial DNA and ESTs, the identification of drug-

induced mutations in HIV, and analysis of mutations in p53 correlated with human cancer.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1: Outline of sequence analysis algorithm using first and second level base calling strategies.

Figs. 2 A-F: Chimpanzee and human chip image comparisons. Magnified digitized false colored red images showing human and chimpanzee *BRCA1* target hybridization patterns to high density oligonucleotide arrays evaluating antisense strands (array size is 1.2 cm x 1.2 cm with 50 micron probe feature sizes). Contrast and brightness parameters were changed in each panel to increase image clarity. Probes designed to detect single nucleotide insertions are not shown for clarity. Nucleotide identities, determined through dideoxysequencing analysis, are given under the respective column, underlined if differing from human, and colored red or blue if correctly or incorrectly identified by level one hybridization analysis, respectively. Several base calls nearby mismatches are difficult to visualize due to limitations in printing technology as well as in the linear range of the human eye for detection of monochromatic color changes. Hybridization patterns of (A) human and (B) chimpanzee (both corresponding to nucleotides 2146-2158 of human cDNA sequence), (C), human and (D), chimpanzee (both corresponding to nucleotides 2125-2137 of human cDNA sequence), and (E), human and (F), chimpanzee (both corresponding to nucleotides 2246-2258 of human cDNA sequence).

Figs. 3 A-G: Primate chip image comparisons. Digitized false colored red images showing hybridization

pattern of *BRCA1* fluorescent targets to high density oligonucleotide arrays evaluating antisense target strands. Magnification of the region (50 micron feature size) corresponding to nucleotide positions 3374-3388 of human *BRCA1* cDNA is given for each species; specific insertion probes are not shown for clarity. The arrangement of sequencing probes is given in Fig. 1B. Nucleotide identities, determined through dideoxysequencing analysis, are given under each column and colored or underlined as described in Fig. 2. Hybridization patterns of (A) human (*Hsa, Homo sapiens*), (B) chimpanzee (*Ptr, Pan troglodytes*), (C), gorilla (*Ggo, Gorilla gorilla*), (D), orangutan (*Ppy, Pongo pygmaeus*), (E), rhesus (*Mmu, Macaca mulatta*), (F), red howler monkey (*Ase, Alouatta seniculus*), and (G), galago (*Gcr, Gala go crassacaudatus*) targets, respectively are shown.

Figs. 4 A-D: Representative chip images of alternative second order tiling schemes for orangutan target sites. Magnified digitized false colored red images showing hybridization pattern of *BRCA1* fluorescent orangutan targets to high density oligonucleotide arrays evaluating sense and antisense target strands. Nucleotide identities, determined through dideoxysequencing analysis, depicting coding strand sequence are given under the respective column and underlined if differing from the human consensus sequence. For the 2731 C->T and 3667 A->G base substitutions relative to human sequence, hybridization to nucleotides 2724-2728 and 3660-3674 using human cDNA numbering are given respectively. Hybridization patterns of orangutan (A) sense target with standard 2731 C tiling, (B), sense target with alternative 2731 T tiling, (C), antisense target with standard 3667 A tiling, (D), antisense target with alternative 3667 G tiling.

Antisense strand hybridization data is given relative to coding strand sequence.

DEFINITIONS

A nucleic acid is a deoxyribonucleotide or ribonucleotide polymer in either single-or double-stranded form, including known analogs of natural nucleotides unless otherwise indicated.

An oligonucleotide is a single-stranded nucleic acid ranging in length from 2 to about 500 bases, and is typically, about 8-40, and more typically, 10-25 bases.

A probe is an oligonucleotide capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. An oligonucleotide probe may include natural (*i.e.* A, G, C, or T) or modified bases (e.g., 7-deazaguanosine, inosine). In addition, the bases in oligonucleotide probe may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, oligonucleotide probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages. See Nielsen et al., *Science* 254, 1497-1500 (1991).

Specific hybridization refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. Stringent conditions are conditions under which a probe will hybridize to its target subsequence, but to no other sequences. Stringent conditions are sequence-dependent and are different in different circumstances. Longer sequences hybridize specifically at higher

temperatures. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength, pH, and nucleic acid concentration) at which 50% of the probes complementary to the target sequence hybridize to the target sequence at equilibrium. (As the target sequences are generally present in excess, at Tm, 50% of the probes are occupied at equilibrium). Typically, stringent conditions include a salt concentration of at least about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.*, 10 to 50 nucleotides). Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM Na phosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridizations.

A perfectly matched probe has a segment perfectly complementary to a particular target sequence. Complementary base pairing means sequence-specific base pairing which includes *e.g.*, Watson-Crick base pairing or other forms of base pairing such as Hoogsteen base pairing. Probes typically have a segment of complementarity of 6-20 nucleotides, and preferably, 10-25 nucleotides. Leading or trailing sequences flanking the segment of complementarity can also be present. The term "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. Although the mismatch(s) may be located anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization of the target sequence. Thus, probes are often designed to have the mismatch located at or near the

center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions.

Polymorphism refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles, each occurring at frequency of greater than 1%, and more preferably greater than 10% or 20% of a selected population. A polymorphic locus may be as small as one base pair.

An array including a pooled probe means that a cell in the array is occupied by pooled mixture of probes. For example, a cell might be occupied by probes ACCCTCCA and ACCCCCCA, in which case, the underline position is described as a pooled position. Although the identity of each probe in the mixture is known, the individual probes in the pool are not separately addressable. Thus, the hybridization signal from a cell is the aggregate of that of the different probes occupying the cell.

The term species variant refers to a gene sequence that is evolutionarily and functionally related between species. For example, in the human genome, the human CD4 gene is the cognate gene to the mouse CD4 gene, since the sequences and structures of these two genes indicate that they are highly homologous and both genes encode a protein which functions in signaling T-cell activation through MHC class II-restricted antigen recognition.

Percentage sequence identity is determined between optimally aligned sequences from computerized implementations of algorithms such as GAP, BESTFIT, FASTA, and TFASTA in the

Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI.

SUMMARY OF THE CLAIMED INVENTION

The invention provides iterative methods of analyzing a target sequence, which represents a variant of a reference sequence. The methods employ an array of probes which includes a probe set comprising probes complementary to the reference sequence. A target nucleic acid is hybridized to the array of probes. The relative hybridization intensities of the probes to the target nucleic acid are then determined. The relative hybridization intensities are used to estimate a sequence of the target nucleic acid. A further array of probes is then provided comprising a probe set comprising probes complementary to the estimated sequence of the target nucleic acid. The target nucleic acid is then hybridized to the further array of probes and the relative hybridization of the probes to the target nucleic acid is determined. The sequence of the target nucleic acid is then reestimate from the relative hybridization intensities of the probes. The cycles of hybridization and estimating the sequence of the target nucleic acid can be reiterated, if desired, until the reestimate sequence of the target nucleic acid is the true sequence of the target nucleic acid.

The methods are particularly useful for analyzing a target nucleic acid that represents a species variant of a known reference sequence. For example, the reference sequence can be from a human and the target sequence from a primate. Typically, the target nucleic acid shows 50-99% sequence identity with the reference sequence. The methods are also particularly useful in situations where a target sequence

differs from a reference sequence by more than one mutation within a probe length.

The methods can readily accommodate a reference sequence of at least 1 or 10 kb long or even a complete or substantially complete human chromosome or genome. A probe set for use in the methods typically includes overlapping probes that are perfectly complementary to and span the reference sequence, and the further array comprises probes that are perfectly complementary to and span the estimate sequence.

In some methods, the array of probes comprises four probe sets. A first probe set comprises a plurality of probes, each probe comprising a segment of at least six nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. Second, third and fourth probe sets, each comprise a corresponding probe for each probe in the first probe set, the probes in the second, third and fourth probe sets being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least six nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the four corresponding probes from the four probe sets. In such methods, the target sequence can be estimated by comparing the relative specific binding of four corresponding probes from the first, second, third and fourth probe sets. A nucleotide in the target nucleic acid is then assigned as the complement of the interrogation position of the probe having the greatest

specific binding. Other nucleotides in the target sequence are assigned by similar comparisons.

The invention also provides methods of analyzing a target nucleic acid comprising the following steps. An array of probes is designed to be complementary to an estimated sequence of the target nucleic acid. The array of probes is hybridized to the target nucleic acid. The target sequence is reestimated from hybridization pattern of the array to the target nucleic acid. The steps are the repeated at least once.

DETAILED DESCRIPTION

1. General

The invention provides improved methods for analyzing variants of a reference sequence using arrays of probes. The methods are particularly useful for target sequences showing substantial variation from a reference sequence, as may be the case where target sequence and reference sequence are from different species. The methods involve designing a primary array of probes based on a known reference sequence. Effectively, the reference sequence serves as a first estimate of sequence of the target nucleic acid. The primary array of probes is hybridized to a target nucleic acid, and the sequence of the target is estimated as well as possible from its hybridization pattern to the primary array. A secondary array of probes is then designed based on the estimated sequence of the target nucleic acid. The target nucleic acid is then hybridized with the secondary array of probes, and the sequence is reestimated from the resulting hybridization pattern. Further cycles of array design and estimation of target sequence can be performed in an iterative

fashion, if desired, until the estimated sequence is constant between successive cycles.

2. Reference Sequences

Reference sequences for polymorphic site identification are often obtained from computer databases such as Genbank, the Stanford Genome Center, The Institute for Genome Research and the Whitehead Institute. The latter databases are available at http://www-genome.wi.mit.edu; http://shgc.stanford.edu and http://ww.tigr.org. Reference sequences are typically from well-characterized organisms, such as human, mouse, *C. elegans*, *Arabidopsis*, *Drosophila*, yeast, *E. coli* or *Bacillus subtilis*. A reference sequence can vary in length from 5 bases to at least 1,000,000 bases. References sequences are often of the order of 100-10,000 bases. The reference sequence can be from expressed or nonexpressed regions of the genome. In some methods, in which RNA samples are used, highly expressed reference sequences are sometimes preferred to avoid the need for RNA amplification. The function of a reference sequence may or may not be known. Reference sequences can also be from episomes such as mitochondrial DNA. Of course, multiple reference sequences can be analyzed independently.

3. Target Nucleic Acid Sample Preparation

Targets can represent allelic, species, induced or other variants of reference sequences. Considerable diversity is possible between reference and target sequence. Target sequences usually show between 50-99%, 80-98%, 90-95% sequence identity. For example, a human reference sequence can be used as the starting point for analysis of primates, such as

gorillas, orangutans, other mammals, reptiles, birds, plants, fungi or bacteria.

The nucleic acid samples hybridized to arrays can be genomic, RNA or cDNA. Nucleic acid samples are usually subject to amplification before application to an array. An individual genomic DNA segment from the same genomic location as a designated reference sequence can be amplified by using primers flanking the reference sequence. Multiple genomic segments corresponding to multiple reference sequences can be prepared by multiplex amplification including primer pairs flanking each reference sequence in the amplification mix. Alternatively, the entire genome can be amplified using random primers (typically hexamers) (see Barrett et al., *Nucleic Acids Research* 23, 3488-3492 (1995)) or by fragmentation and reassembly (see, e.g., Stemmer et al., *Gene* 164, 49-53 (1995)). Nucleic acids can also be amplified by cloning into vectors and propagating the vectors in a suitable organism. YACs, BACs and HACs are useful for cloning large segments of genomic DNA.

Genomic DNA can be obtained from virtually any tissue source (other than pure red blood cells). For example, convenient tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair.

RNA samples are also often subject to amplification. In this case amplification is typically preceded by reverse transcription. Amplification of all expressed mRNA can be performed as described by commonly owned WO 96/14839 and WO 97/01603. In some methods, in which arrays are designed to tile highly expressed sequences, amplification of RNA is unnecessary. The choice of tissue from which the sample is obtained affects the relative and absolute levels of different

RNA transcripts in the sample. For example, cytochromes P450 are expressed at high levels in the liver.

4. Methods of amplification

The PCR method of amplification is described in *PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (eds. McPherson et al., IRL Press, Oxford); and U.S. Patent 4,683,202 (each of which is incorporated by reference for all purposes). Nucleic acids in a target sample are usually labelled in the course of amplification by inclusion of one or more labelled nucleotides in the amplification mix. Labels can also be attached to amplification products after amplification e.g., by end-labelling. The amplification product can be RNA or DNA depending on the enzyme and substrates used in the amplification reaction.

Other suitable amplification methods include the ligase chain reaction (LCR) (see Wu and Wallace, *Genomics* 4, 560 (1989), Landegren et al., *Science* 241, 1077 (1988), transcription amplification (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989)), and self-sustained sequence replication (Guatelli et al., *Proc. Nat. Acad. Sci. USA*, 87, 1874 (1990)) and nucleic acid based sequence amplification (NASBA). The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

5. Probe Arrays

An array of probes contain at least a first set of probes that are complementary to a reference sequence (or regions of interest therein). Typically, the probes tile the reference sequence. Tiling means that the probe set contains overlapping probes which are complementary to and span a region of interest in the reference sequence. For example, a probe set might contain a ladder of probes, each of which differs from its predecessor in the omission of a 5' base and the acquisition of an additional 3' base. The probes in a probe set may or may not be the same length. The number of probes can vary widely from about 5, 10, 20, 50, 100, 1000, to 10,000 or 100,000. Typically, the arrays do not contain every possible probe sequence of a given length.

Often tiling arrays have four probe sets, as described in WO 95/11995. The first probe set comprises a plurality of probes exhibiting perfect complementarily with a reference sequence, as described above. Each probe in the first probe set has an interrogation position that corresponds to a nucleotide in the reference sequence. That is, the interrogation position is aligned with the corresponding nucleotide in the reference sequence, when the probe and reference sequence are aligned to maximize complementarily between the two. For each probe in the first set, there are three corresponding probes from three additional probe sets. Thus, there are four probes corresponding to each nucleotide in the reference sequence. The probes from the three additional probe sets are identical to the corresponding probe from the first probe set except at the interrogation position, which occurs in the same position in each of the four corresponding probes from the four probe sets, and is occupied by a different nucleotide in the four probe sets.

A substrate bearing the four probe sets is hybridized to a labelled target sequence, which shows substantial sequence similarity with the reference sequence, but which may differ due to e.g., species variations. The amount of label bound to probes is measured. Analysis of the pattern of label revealed the nature and position of differences between the target and reference sequence. For example, comparison of the intensities of four corresponding probes reveals the identity of a corresponding nucleotide in the target sequences aligned with the interrogation position of the probes. The corresponding nucleotide is the complement of the nucleotide occupying the interrogation position of the probe showing the highest intensity. The comparison can be performed between successive columns of four corresponding probes to determine the identity of successive nucleotides in the target sequence.

In many instances of comparing four corresponding probes, one of the four probes clearly has a significantly higher signal than the other three, and the identity of the base in the target sequence aligned with the interrogation position of the probes can be called with substantial certainty. However, in some instances, two or more probes may show similar but not identical signals. In these instances, one can simply score the position as ambiguous. Alternatively, can still call a base from the probe that has the higher signal but must recognize a significant possibility of error. In general, if the ratio of signals of two probes is less than 1.2, a base call has a significant possibility of error. Ambiguous positions are most frequently due to closely spaced multiple points of variation between target and reference sequence (i.e., within a probe length). Ambiguities

can also arise due to low hybridization intensity because of base composition effects.

A secondary array of probes is constructed based on the same principles as the first array, except that the first probe set is tiled based on the newly estimated sequence rather than the original reference sequence. In general, the estimated sequence includes the best estimate of base present at positions of ambiguity as noted above. If there is equal probability of two or more bases occupying a particular position in the estimated sequence, one can arbitrarily decide to include one of the bases, provide alternate tilings corresponding to the different possible bases, or include multiple pooled bases at the position. The secondary array typically has second, third and fourth probe sets designed according to the same principles as in the primary array.

The secondary array is hybridized to the same target nucleic acid as was the primary array. Bases in the target sequence are called using the same principles as described above by comparison of probe intensities to give rise to a reestimated target sequence.

The process can be repeated through further iterations, if desired. Further iteration is desirable if the estimated sequence contains a substantial number of positions, which have been estimated with a low degree of confidence (e.g., from a comparison of probe intensities differing by a factor of less than 1.2). After sufficient iterations, the estimated sequence from one cycle should converge with that from the subsequent cycle. In some instances, positions of ambiguities may remain through many cycles. These positions may be due to effects such as heterozygosity, and should be checked by other means (e.g., conventional dideoxy sequencing

or de novo sequencing by hybridization to a complete array of probes a given length).

Many variations in array design and analysis are possible, as described for example in WO 95/11995; EP 717,113; WO 97/29212. Optionally, arrays tile both strands of a reference sequence. Both strands are tiled separately using the same principles described above, and the hybridization patterns of the two tilings are analyzed separately. Typically, the hybridization patterns of the two strands indicates the same results (i.e., location and/or nature of variation between target sequence and reference sequence). Occasionally, there may be an apparent inconsistency between the hybridization patterns of the two strands due to, for example, base-composition effects on hybridization intensities. Combination of results from the two strands increases the probability of correct base calling and can decrease the number of iterations required to determine the correct base sequence of a target.

In a further variation, duplicate arrays are synthesized to allow analysis of hybridization between target sequence and probes under conditions of high and low stringency. Although high stringency is generally most useful, there are some regions of target sequence where the absolute hybridization intensity is low due to base composition effects, which yield base calls with a higher degree of confidence under conditions of low stringency. Statistical combination of base calls from conditions of high and low stringency can increase the overall probability of correct base calling.

6. Synthesis and Scanning of Probe Arrays

Arrays of probe immobilized on supports can be synthesized by various methods. A preferred methods is VLSIPS™ (see Fodor et al., US 5,143,854; EP 476,014, Fodor et al., 1993, *Nature* 364, 555-556; McGall et al., USSN 08/445,332), which entails the use of light to direct the synthesis of oligonucleotide probes in high-density, miniaturized arrays (sometimes known as chips). Algorithms for design of masks to reduce the number of synthesis cycles are described by Hubbel et al., US 5,571,639 and US 5,593,839. Arrays can also be synthesized in a combinatorial fashion by delivering monomers to cells of a support by mechanically constrained flowpaths. See Winkler et al., EP 624,059. Arrays can also be synthesized by spotting monomers reagents on to a support using an ink jet printer. *See id.*; Pease et al., EP 728,520.

After hybridization of control and target samples to an array containing one or more probe sets as described above and optional washing to remove unbound and nonspecifically bound probe, the hybridization intensity for the respective samples is determined for each probe in the array. For fluorescent labels, hybridization intensity can be determined by, for example, a scanning confocal microscope in photon counting mode. Appropriate scanning devices are described by e.g., Trulson et al., US 5,578,832; Stern et al., US 5,631,734.

7. Large-Scale Resequencing

The methods described above can be used for comparative analysis of whole genomes or substantial portions thereof. To illustrate, about 300 chips at 1 Mb/chip are required to sequence 10% of a mammalian genome (i.e., all the genes and a substantial amount of their surrounding sequence).

If 40 chips are synthesized on a common waver using a single mask, then only 8 mask designs are required per iteration. If 10 iterations are required, then only 80 mask designs and a total of 3000 chips are made.

Although an entire genome can be hybridized to a chip in a single experiment, it is often more useful to hybridize pools of cloned sequence representing ~ 1 Mb at a time. This can be done in the following way. A minimal overlapping set of physical clones is first obtained. For example, random bacterial artificial chromosome clones are generated, and ordered by hybridization or conventional methods. If necessary, regions mapping to related positions in the genome are determined. E.g., pools of clones are hybridized to an array of mapped markers. Pools of clones are then generated for hybridization (e.g., 300 pools if the resequencing capacity is 1 Mb/chip and 300 chip designs are used to analyze 1/10th a mammalian genome).

8. Applications

Some of the benefits of resequencing related genomes are:

1) Correction of sequencing errors. These are often corrected by comparative analysis. For example, if an open reading frame in one genome is frameshifted in a second closely related genome, a sequencing error is usually the cause of the difference. Any sequence differences detected can be verified in the reference genome by simply checking the primary sequencing trace data, or by further analysis.

2) Identification of promoter sequences and genes. Functionally important elements tend to be conserved. Sometimes, functional elements that are difficult to identify by direct sequence analysis (such as small exons or regulatory

sequences) are revealed by identifying relatively short segments that are tightly conserved between genomes.

3) Analysis of sequences differences between differences species allows correlation between form and function. For example, the sequence of chimpanzee and human differ by ~1% overall. Further, the present methods allow comparison of a range of primate sequences, to see which sequences have evolved the most rapidly and which are highly conserved.

It will be apparent from the above that the invention includes a general concept which can be expressed concisely as follows. The invention entails the use of iterative cycles of designing an array of probes to be complementary to an estimated sequence of a target nucleic acid, and using the hybridization pattern of the array to the target nucleic acid sequence to determine a more accurate reestimated target sequence.

EXAMPLES

We examined the use of high density oligonucleotide arrays (DNA chips) to obtain sequence information from homologous genes in closely related species. Orthologs of the human *BRCA1* gene exon 11, all approximately 3.4 kb in length and ranging from 98.2% to 83.5% nucleotide identity, were subjected to hybridization-based and conventional dideoxysequencing analysis. Based upon dideoxysequencing results, retrospective guidelines for identifying high fidelity hybridization-based sequence calls were formulated. Prospective application using these rules yielded base calling with at least 98.8% accuracy over orthologous sequence tracts shown to have approximately 99% identity. For higher primate sequences with greater than 97% nucleotide identity, base

calling was made with at least 99.91% accuracy covering a minimum of 97% of the sequence. Using a second tier confirmatory hybridization chip strategy shown in several cases to confirm the identity of predicted sequence changes, the complete sequence of the chimpanzee, gorilla, and orangutan orthologs was deduced solely through hybridization-based methodologies. Analysis of less highly conserved orthologs identified conserved nucleotide tracts of at least 15-nt in length, which are useful in design of further arrays of hybridization probes or in the design of primers.

I. Methods

(1) PCR from genomic DNA and transcription

PCR reactions were performed on genomic samples using the EXPAND™ Long Range PCR Kit (Boehringer Mannheim) with intronic primers 11PIF 5'-CCTTGTTATTTTTTGTATATTTTCAG-3' and 11PIR 5'-CAAAAACCTGGTTCCAATAC-3', directly overlapping the underlined 5'-AG acceptor and 3'-GT donor splice sites. PCR reactions, using the templates generated from the 11PIF and 11PIR primer set, were performed with primers 11PIFT3 5'-ATTAACCCTCACTAAAGGGACCTTGTTATTTTTTGTATATTTTCAG-.3' and 11PIRT7 5'-TAATACGACTCACTATAGGGACAAAAACCTGGTTCCAA TAC-3' containing T3 and T7 RNA polymerase promoter sequences respectively. *In vitro* transcription reactions from these templates were performed in 10μ1 reaction volumes using T3 RNA polymerase transcription buffer (Promega), 0.7 mM of ATP, CIP, GTP, and UTP, 10 mM DTT, 0.15 mM biotin-16-UTP (Boehringer Mannheim) and 10 U T3 or T7 RNA polymerase as indicated.

(2) Target preparation and analysis

Test sample transcription product was diluted to a final concentration of 100 nM in a 25 μl solution of 30 mM $MgCl_2$. The reaction was incubated at 94 degrees C for 60 minutes to hydrolyze target into fragments ranging from (50-100)-nt in length and subsequently diluted 1/100 into a 300 μl volume of hybridization buffer (3 M TMAC (tetramethylammonium chloride), 1X TE pH 7.4, 0.005% Triton X-100, 1 nM 5'-fluorescein-labelled control oligonucleotide 5'CGGTACCATCTTGAC-3').[10] The control oligonucleotide hybridizes to specific surface probes aiding in image alignment. Target was hybridized with the appropriate sense or antisense strand reading array in a 250 μl volume for 4 hours at 35 degrees C. The array surface was washed with 5 ml of wash buffer (6X SSPE, 0.005% Triton X-100) and stained with phycoerythrin-streptavidin conjugate (Molecular Probes) (2 μ/ml in wash buffer containing 2 mg/ml acetylated BSA (GIBCO BRL)) for 20 minutes at room temperature. Each array was washed with 5 ml of wash buffer and imaged using a 488 nm argon laser equipped with a scanning confocal microscope (GeneChip Scanner, Affymetrix). Fluorescent hybridization signals were detected by a photomultiplier tube using a 560 nm longpass emission filter.

(3) Oligonucleotide array synthesis and design

The synthesis and design of the oligonucleotide array has been described previously[9]. Briefly, DNA phosphoramidites bearing 5'-photolabile protecting groups are coupled to a derivatized glass surface using modified DNA synthesis protocols. Spatially addressable oligonucleotide synthesis is obtained through photolithographic techniques with selective oligonucleotide photodeprotection for each

coupling cycle. Thirty identical high density array chips containing over 48,000 oligonucleotides were simultaneously produced in a single 8 hour synthesis.

(4) Chip hybridization experiments

GeneChip Software (Affymetrix) created digitized fluorescence images by converting photomultiplier tube output signal into proportional spatially addressed pixel values. The probe intensity is calculated from the mean of the non-outlier photon counts for each feature (i.e. per probe). Background corrected fluorescent hybridization signal to each probe was extracted from test images using AVI Software (Affymetrix) and imported into ViewSeq Software (Affymetrix) which quantitates ratios of fluorescent target hybridization signal to each set of 8 oligonucleotide probes (4 per strand) interrogating each nucleotide. Data from 4 sets of experiments reading both target strands were averaged to produce a composite file.

(5) Dideoxysequencing analysis

Template PCR products were purified using the Wizard DNA Purification Kit (Promega). Conventional fluorescent dye terminator 3 pass dideoxysequencing analysis was performed using the AB1377 System. Human *BRCA1* exon 11 primers were used for first pass sequencing of all templates, except the canine ortholog of known sequence[18]. Sequence gaps were filled by a primer walking strategy. This sequencing and template generation strategy is not sensitive towards detecting all possible heterozygous single nucleotide polymorphisms; however, it is quite sensitive to detection of heterozygous sequences causing chain length differences in dideoxysequencing products. Nevertheless, in cases of

heterozygous base substitutions the identity of one allele is reported. A nested amplicon within the flying lemur template was generated using Amplitaq GOLD (Perkin Elmer) and the manufacturers protocols to clarify a suspected heterozygous sequence. PCR product was subcloned using Zero Blunt Cloning Kit (Invitrogen) and inserts from individual colonies were sequenced. A heterozygous in-frame 3 base pair deletion was found in flying lemur target which aligns with bases 2192-2194 of human *BRCA1* cDNA sequence and results in the removal of a single serine from a tract of three serine residues. Orthologous sequences were submitted to GenBank and assigned the following accession numbers: AF019075 (chimpanzee), AFOl 9076 (gorilla), AF019077 (orangutan), AFOI 9078 (rhesus), AF019079 (red howler monkey), AF019080 (galago), and AFO190S1 (flying lemur).

(6) Dideoxysequencing data alignment

The multiple alignments of nucleotide sequences were computed using a map program that utilizes a global, optimal alignment algorithm and fixed penalty for long gaps[19]. We used the following parameters for computing optimal alignment: match score = 10, mismatch (DNA alignment) = -5, maximum length of gap to be penalized =10, gap opening penalty = 50, gap extension penalty 5.

II. Results

High density arrays have been used to screen the 3.43-kb exon 11 of the human hereditary breast and ovarian cancer *BRCA1* gene[13] for all possible heterozygous polymorphisms and mutations[9]. Four 20-nt long probes, substituted in the central position with one of the four nucleotides, interrogate the identity of each human *BRCA1* exon 11 nucleotide. In the

algorithm proposed here (Fig. 1), level one analysis quantitates the ratios of fluorescent target hybridization signal for eight probes (four per sense and antisense strands, respectively) querying each nucleotide position. If the ratios between the brightest and next brightest probe signals in each set is greater than 1.2, the identity of the brightest signal is assigned to the target nucleotide, using human sequence numbering. If the brightest probe signal in each set is less than or equal to a factor of 1.2 of the next brightest, an IUPAC ambiguity designation is assigned. (A similar algorithm analyzed regions of the HIV protease gene[7] and human mitochondrial genome[3] with >99.9% accuracy.) For human BRCA1 exon 11 target, 1,363 nucleotide positions had single nucleotide mismatch specificity ratios (the ratio between the two highest probe signals within each averaged composite set of four) greater than 9.0, 1,346 positions had ratios between 5.0 and 9.0, 708 positions had ratios between 2.0 and 5.0, 5 positions had ratios between 1.2 and 2.0, and 4 positions had ratios less than 1.2 (giving an ambiguous level one base call).

Orthologous *BRCA1* exon 11 sequences were subjected to dideoxysequence analysis (Methods, Table 1) and independently assayed on human *BRCA1* sequence based chips. The majority of the differences between humans and other species were single nucleotide substitutions. We found 3 common hybridization patterns associated with single nucleotide substitutions flanked by over 10-nt of homology on each side (Fig. 2, Table 2). For many single base changes, level one hybridization data can accurately identify nucleotide sequences (Fig 2B). As expected, however, single nucleotide substitutions destabilize other "perfect match" probes resulting in a footprint or loss of intensity for neighboring

probes[3-9]. Depending on sequence context, this can lead to an adjacent misidentified base (Fig 2D) or several misidentified or unscorable bases associated with diminished flanking hybridization signals (Fig 2F). In all these cases, the "perfect match" substitution probe corresponding to the nucleotide change is substantially brighter than the surrounding probes.

Human, chimpanzee, gorilla, and orangutan targets with identical sequence tracts showed similar hybridization patterns (Figs. 3 A-D). In this example, a single nucleotide substitution between rhesus and human targets is correctly identified by level one analysis; however, the 3'-adjacent nucleotide is incorrectly assigned (Fig. 3E). Level one hybridization data identifies two red howler monkey nucleotide substitutions, but cannot accurately read adjacent sequences (Fig. 3F). Galago target contained 3 closely spaced nucleotide substitutions causing diminished hybridization signals and lower fidelity nucleotide assignments (Fig. 3G).

We determined the accuracy of level one sequence information from the least (dog) and most (chimpanzee) highly conserved targets by referring to dideoxysequencing data. Upon inspecting level one dog sequence, it was evident that base calling was poor quality in regions of predicted multiple substitutions. Furthermore, it was apparent that the most accurate level one base calls occurred in sequence tracts predicted identical to human reference sequence. Therefore, such tracts ranging from 4 to 8 nucleotides in length were systematically evaluated for base-calling fidelity. In addition, predicted single nucleotide substitutions flanked by these tracts were included in this evaluation since the array has the capacity to correctly identify them (Fig. 2B). Although all these nucleotide tracts gave similar base-calling

accuracy, a 7 nucleotide window size gave the best signal to noise ratio in chimpanzee. We prospectively applied this algorithm to level one sequence data from the remaining species (Table 1, Fig. 1). Proposed level one nucleotide sequences were first aligned with reference human sequence. Second, all 7 nucleotide long tracts with complete identity to human reference sequence as well as single nucleotide substitutions flanked by such tracts were identified. Tracts fulfilling these criteria were defined as level two sequence information.

To clarify this algorithm, consider Figure 2. The data in Figure 1B was judged acceptable for level two since only one nucleotide position, flanked on both 5'- and 3'- ends by 7 nucleotides (6 shown) of predicted homology, was assigned to differ from human. Both the misidentified base and the correctly identified variant nucleotide in Figure 2D were disregarded since 2 adjacent nucleotide positions were predicted to differ from human. The sequence on either side of these 2 bases could still be accepted as level two. The central 5' CTAACT-3' level one sequence in Figure 2F was not accepted for level two since it was in a cluster of predicted changes. Flanking sequences were retained since they were parts of predicted tracts of identity.

Level one and two analysis call the complete human *BRCA1* exon 11 sequence with 99.88% accuracy (Table 1). Both schemes agree since level one miscalls were not clustered within 7 nucleotides of one another. Of the 4 miscalls, 3 were due to insufficient signal and one was due to crosshybridization to another substitution probe. These miscalls and ambiguities were reproducible, suggesting that two color hybridization experiments would be useful in eliminating this source of error as well as in identifying

heterozygous sequence changes, likely found in targets derived from different individuals.[3,9]

All 28 chimpanzee, 31 gorilla, and 63 orangutan nucleotide substitutions compared to human (determined by dideoxysequencing analysis) were correctly identified in level one analysis. For 15 chimpanzee, 22 gorilla, and 36 orangutan base substitutions, the flanking nucleotides were also correctly identified. A total of 17/28 chimpanzee, 14/23 gorilla, and 49/55 orangutan level one miscalls (falsely identified nucleotides compared to dideoxysequencing data) were found within 7 nucleotides of a base substitution. All chimpanzee, gorilla, and orangutan miscalls were found in regions of low signal (<100 counts) or specificity (<1.4X intensity discrimination between two highest intensity probes) in higher primate targets.

When complete sequencing is desired, a second order tiling scheme, with probes designed to match anticipated base substitutions in the level one data based upon single nucleotide mismatch probe hybridization signals, clarifies most or all ambiguities. We unambiguously confirmed two predicted base changes present in chimpanzee, gorilla, and orangutan targets (2731T and 3667G, using human cDNA numbering) using a second order tiling scheme (Fig. 4). All chimpanzee, gorilla, and orangutan miscalls made adjacent to base substitutions can be clarified using second order tiling schemes since the sequence accuracy was at least 99.88% when the tiling pattern matches the target sequence.

More poorly conserved rhesus, red howler monkey, galago, and dog orthologs provided less level two quality hybridization data (Table 1). This was primarily caused by an increased number of closely spaced nucleotide substitutions along with insertions and deletions. Of the 26 level-two red

howler target miscalls, 23 were found nearby an almost exact 21-bp target duplication while 3 were due to a 3 base pair deletion.

Completion of the Human Genome Project allows use of DNA chips for rapid genome-wide determination of non-human primate sequences[14]. This approach is particularly powerful when scanning for conserved sequence tracts, important for phylogenetic footprinting of promoter regions[2]. A pair of high density oligonucleotide arrays containing 20 μm feature sizes, which can be currently be manufactured using photolithographic processes and imaged with commercially available instrumentation, can analyze over 100-kb of sequence.

Bibliography

1. Li et al., *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).

2. Tagle et al., *J. Mol. Biol.* 203:439-455 (1988).

3. Chee et al., *Science* 274: 610-614 (1996).

4. Ormanac et al., *Science* 260:1649-1652 (1993).

5. Fodor et al., *Science* 251:767-773 (1991).

6. Cronin et al., *Hum. Mut.* 7: 244-255 (1996).

7. Kozal et al., *Nature Med.* 2:753-759 (1996).

8. Yershov et al., *Proc. Natl. Acad. Sci. USA* 93:4913-4918 (1996).

9. Hacia et al., *Nature Gen.* 14:441-447 (1996).

10. Lockhart et al., *Nature Biotech.* 14:1675-1680 (1996).

11. Shoemaker et al., *Nature Genet.* 14:450456 (1996).

12. Milner et al. *Nature Biotech.* 15:537-541 (1997).

13. Miki et al., *Science* 266:66-71 (1994).

14. McConkey et al., *Trends in Genet.* 13: 350-351 (1997).

15. Kruglyak, *Nature Genet.* 17:21-24 (1997).

16. Hoheisel, *Nucleic Acids Res.* 24:430-432 (1996).

17. Broude, *Proc. Natl. Acad. Sci. USA.* 91: 3072-3076 (1994).

18. Szabo et al., *Hum. Mol. Genet.* 5:1289-1.298 (1996).

19. Huang, *Appl. Biosci.* 10:227-235 (1994).

All publications and patent applications cited above are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent application were specifically and individually indicated to be so incorporated by reference. Although the present invention has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims.

Table 1 • Evaluation of Hybridization Based Sequence Calls for *BRCA1* Exon 11

Species	Nt Identity[a]	Bases Analyzed[b]	Correct[c]	Incorrect[d]	Level 2 Identity[e]
Homo sapiens	100	3426 (100)	3422 (99.88)	4 (0.12)	100
Pan troglodytes	99.2	3379 (98.63)	3376 (99.91)	3 (0.09)	99.52
Gorilla gorilla	99.1	3387 (98.86)	3384 (99.91)	3 (0.09)	99.32
Pongo pygmaeus	98.2	3321 (96.94)	3320 (99.97)	1 (0.03)	98.83
Macaca mulatta	95.9	2990 (87.27)	2985 (99.83)	5 (0.17)	98.56
Alouatta seniculus	92.6	2364 (69.00)	2338 (98.90)	26 (1.10)	98.65
Galago crassicaudatus	85.0	1296 (37.83)	1290 (99.54)	6 (0.46)	99.07
Cynocephalus variegatus	87.2	1621 (47.31)	1616 (99.69)	5 (0.31)	98.95
Canis familiaris	83.5	1237 (36.11)	1233 (99.68)	4 (0.32)	98.79

[a]Per cent nucleotide identity of complete target relative to human based on dideoxysequencing analysis.

[b]Nucelotides located in tracts of at least seven bases of identity to human, or single nucleotide substitutions flanked by at least seven bases of identity. Percentage of the data that met level two criteria is shown in parenthesis. [c]Correct level two calls with percentage of analyzed bases shown in parenthesis. [d]Incorrect level two cells with miscall percentage of analyzed bases shown in parenthesis.

[e]Per cent nucleotide identity of level two data relative to human target.

What is claimed is:

1. A method of analyzing a target nucleic acid, comprising:

(a) providing an array of probes comprising a probe set comprising probes complementary to a reference sequence;

(b) hybridizing the target nucleic acid to the array of probes;

(c) determining the relative hybridization of the probes to the target nucleic acid,

(d) estimating the sequence of the target nucleic acid from the relative hybridization of the probes;

(e) providing a further array of probes comprising a probe set comprising probes complementary to the estimated sequence of the target nucleic acid;

(f) hybridizing the target nucleic acid to the further array of probes;

(g) determining the relative hybridization of the probes to the target nucleic acid;

(h) reestimating the sequence of the target nucleic acid from the relative hybridization of the probes.

2. The method of claim 1, further comprising repeating steps (e)-(h) as necessary until the reestimated sequence of the target nucleic acid is the true sequence of the target nucleic acid.

3. The method of claim 1, wherein the target nucleic acid is a species variant of the reference sequence.

4. The method of claim 1, wherein the reference sequence is from a human and the target nucleic acid is from a primate.

5. The method of claim 1, wherein the target nucleic acid shows 50-99% sequence identity with the reference sequence.

6. The method of claim 1, wherein the target nucleic acid shows 80-95% sequence identity with the reference sequence.

7. The method of claim 1, wherein the reference sequence is at least 1000 nucleotides long, the array comprises a probe set comprising overlapping probes that are perfectly complementary to and span the reference sequence, and the further array comprises probes that are perfectly complementary to and span the estimated sequence.

8. The method of claim 1, wherein an estimated sequence of the target nucleic acid includes a position of ambiguity and the probe set showing perfect complementarity to the estimated sequence includes a probe having including a pooled nucleotide aligned with the position of ambiguity in the target sequence.

9. The method of claim 1, wherein the reference sequence is at least 10 kb.

10. The method of claim 1, wherein the reference sequence is at least 1000 kb.

11. The method of claim 1, wherein the reference sequence includes at least 90% of the human genome.

12. The method of claim 1, wherein the array of probes comprises:

(1) a first probe set comprising a plurality of probes, each probe comprising a segment of at least six nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence,

(2) second, third and fourth probe sets, each comprising a corresponding probe for each probe in the first probe set, the probes in the second, third and fourth probe sets being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least six nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the four corresponding probes from the four probe sets.

13. The method of claim 12, wherein the sequence of the target nucleic acid is estimated by:

(a) comparing the relative specific binding of four corresponding probes from the first, second, third and fourth probe sets;

(b) assigning a nucleotide in the sequence of the target nucleic acid as the complement of the interrogation position of the probe having the greatest specific binding;

(c) repeating (a) and (b) until each nucleotide of interest in the sequence of the target nucleic acid has been estimated.

14. The method of claim 1, wherein the sequence of the target nucleic acid differs from the reference by at least two positions within a probe length.

15. The method of claim 1, wherein the estimating or reestimating comprises determining from the relative hybridization of the probes positions of single nucleotide substitution in the target nucleic acid relative to the reference sequence from, which positions are flanked on both sides by segments of at least seven contiguous nucleotides of identity between the target nucleic acid relative to the reference sequence and including the bases occupying the positions in the reestimated sequence.

16. The method of claim 15, wherein the reestimated or estimate sequence of the target nucleic acid is the same as the estimated sequence or reference sequence respectively except at the bases occupying the positions of single nucleotide substitution.

17. A method of analyzing a target nucleic acid, comprising:

(a) designing an array of probes to be complementary to an estimated sequence of the target nucleic acid,

(b) hybridizing the array of probes to the target nucleic acid;

(c) determining a reestimated sequence of the target nucleic acid from the hybridization pattern of the array to the target nucleic acid sequence to; and

(d) repeating (a)-(c) at least once.

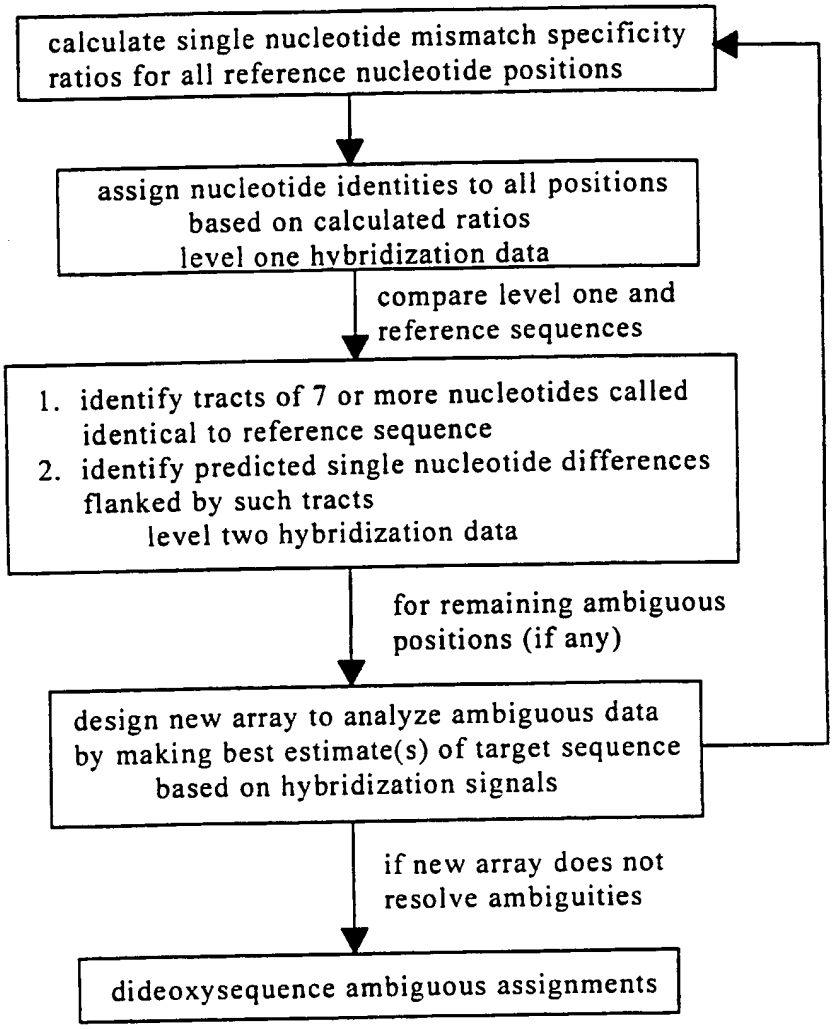
calculate single nucleotide mismatch specificity ratios for all reference nucleotide positions
assign nucleotide identities to all positions based on calculated ratios level one hybridization data
compare level one and reference sequences
1. identify tracts of 7 or more nucleotides called identical to reference sequence
2. identify predicted single nucleotide differences flanked by such tracts
level two hybridization data
for remaining ambiguous positions (if any)
design new array to analyze ambiguous data by making best estimate(s) of target sequence based on hybridization signals
if new array does not resolve ambiguities
dideoxysequence ambiguous assignments

FIG. 1

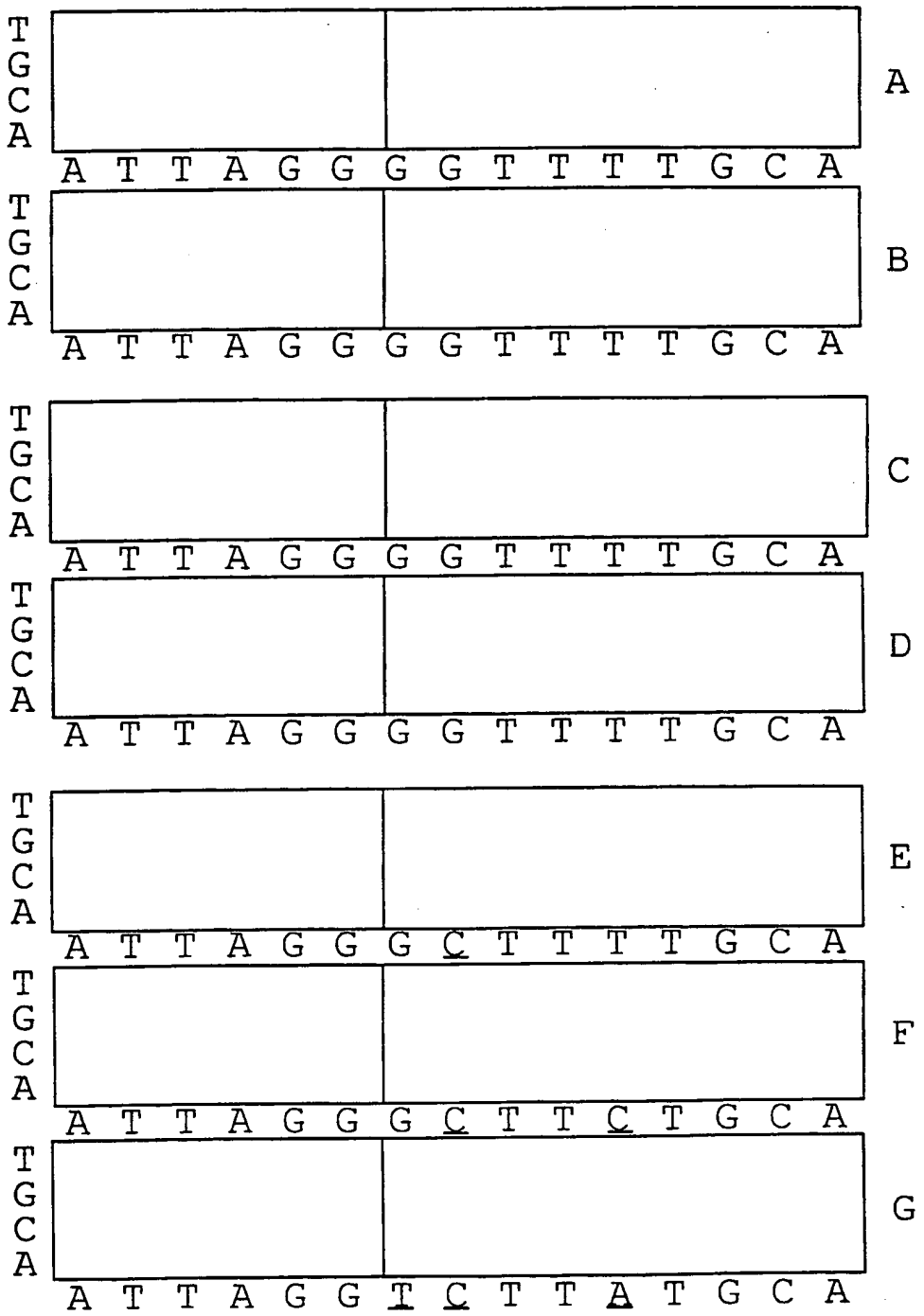

FIG. 3

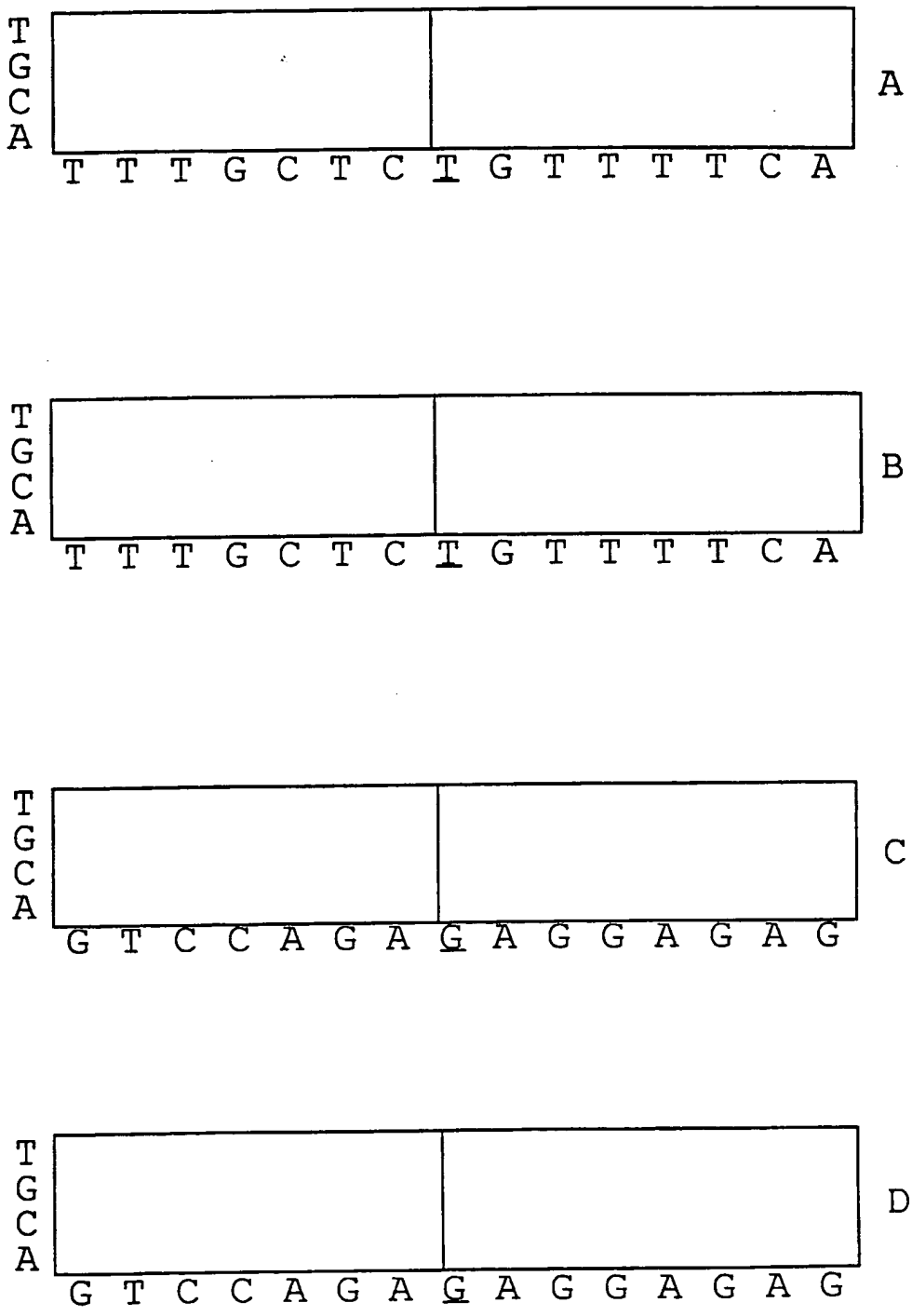
TGCA
T T T G C T C T G T T T T C A
A
TGCA
T T T G C T C T G T T T T C A
B
TGCA
G T C C A G A G A G G A G A G
C
TGCA
G T C C A G A G A G G A G A G
D

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/05438

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; C07H 21/04
US CL : 435/6; 536/23.1, 24.31

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/23.1, 24.31

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DIALOG, APS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,698,391 A (COOK et al.) 16 December 1997, see especially the abstract and columns 3-6.	1-17
X	US 5,683,881 A (SKIENA) 04 November 1997, see especially the abstract and columns 3-5.	1-17

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 08 MAY 1998	Date of mailing of the international search report 18 JUN 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer WILLIAM SANDALS Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*